Position paper d'ANITI sur les systèmes d'IA générative

ANITI

Université de Toulouse

Préambule : *Le déploiement récent et sans précédent des systèmes d'IA générative a donné lieu à de très nombreux travaux et à une intense couverture médiatique. ANITI a mis en place un groupe de travail en vue notamment de préciser son programme de recherche sur* ***l'IA générative*** *dans le cadre de sa stratégie sur les systèmes critiques, en particulier dans les transports et l'industrie, ainsi que sur l'IA de confiance socialement acceptable et responsable.* ***Le groupe de travail a produit ce document qui synthétise la position d'ANITI partagée par ses porteurs de Chaires.***

L'IA générative a permis le développement d'une technologie qui est impressionnante au moins par le nombre d'utilisateurs qu'elle a rencontrés. Certaines de ses utilisations rendent les scientifiques et observateurs de la société parfois enthousiastes, parfois perplexes, ou inquiets. Jamais une avancée en IA n'avait à ce point captivé le grand public. Les systèmes d'IA générative démontrent des aptitudes **non anticipées et encore non expliquées,** mais aussi bien trop fragiles pour leur confier des tâches essentielles. Ils renvoient à des questions scientifiques ouvertes pour expliquer leurs performances et cerner leurs limitations, ainsi qu'à des questions philosophiques profondes autour de la manière dont nous définissons l'intelligence.

L'écrasante majorité de ces systèmes d'IA générative est développée par un petit nombre d'acteurs industriels (les GAFA et leurs filiales). Ceci soulève une légitime préoccupation **d'indépendance nationale et de souveraineté**. Mais ceci a également une incidence considérable sur la nature des développements et l'évolution globale du domaine : la logique industrielle privilégie naturellement le savoir-faire opérationnel au détriment de la compréhension scientifique de ce qui est mis en œuvre. Sans une telle compréhension, ces systèmes restent opaques, non contrôlables, et leur régulation par la société très difficile. **Or le développement d'une IA de confiance et responsable, auquel ANITI est consacré, passe par l'appropriation scientifique de tels systèmes.**

En dehors de réactions extrêmes qui consistent à nier le caractère transformateur de cette technologie, ou bien à prophétiser son rôle dans l'asservissement, voire l'extinction de l'espèce humaine, c'est l'incertitude qui domine, parfois teintée d'inquiétude. Cette incertitude est justifiée. Le langage par exemple est à la base des structures sociales et des civilisations. Il imprègne toutes les activités humaines. Par conséquent, les grands modèles génératifs de langue, sur lesquels nous nous focalisons dans la suite du présent texte pour fixer les idées, ont un potentiel de transformation considérable.

Certaines de ces transformations présentent des risques éthiques et sociaux connus et avérés, par exemple de dépendance accrue aux machines, de manipulation sociale à grande échelle, d'exacerbation des discriminations, ou d'augmentation de la capacité de nuisance d'acteurs malintentionnés. D'autres transformations, cependant, ne sont pas intrinsèquement négatives. Les grands modèles de langage vont bouleverser notre manière de travailler, d'étudier, de nous informer, de communiquer, et de nous comporter les uns envers les autres.

Ces bouleversements auront des conséquences, positives et négatives, qu'il nous faut anticiper, accompagner, et permettre à la société de maîtriser et réguler. On peut imaginer que les grands modèles de langue entraîneront une transformation du marché du travail, en revalorisant certaines compétences et en en dévalorisant d'autres. Un bouleversement pourrait se manifester aussi dans le secteur de l'éducation, en orientant les choix et méthodes d'apprentissage pour préparer au mieux les jeunes générations à un monde où la gestion du langage sera parfois déléguée aux machines. Les grands modèles de langue rendront plus aisée la production d'information textuelle, dont l'authentification et l'évaluation de la fiabilité seront névralgiques. Il est peu probable que nous puissions naviguer dans cette marée d'informations sans l'aide d'outils d'intelligence artificielle. Allons-nous vers un avenir où les humains ne seront plus les seuls acteurs du langage ? Il nous faut définir comment pallier les risques de ces transformations et comment en tirer le meilleur parti. Il est impératif de solliciter les sciences sociales et comportementales pour définir ensemble les valeurs et les normes sociales que nous souhaitons préserver, mais aussi pour anticiper celles qui émergeront naturellement dans une société où le langage sera devenu largement artificiel. Les instituts interdisciplinaires d'intelligence artificielle (3IAs), dont ANITI, offrent un cadre unique de collaboration entre sciences humaines et sciences du numérique pour accompagner ces transformations.

Comme pour toute technologie émergente à fort potentiel de transformation, il convient en premier lieu de bien saisir les aspects scientifiques et mathématiques derrière l'IA générative. Il faut notamment comprendre les limites des systèmes d'apprentissage automatique lorsqu'ils sont de très grande dimension et basés sur des corpus importants de données. S'agissant de textes, les grands modèles génératifs actuels les plus utilisés (ils ont des centaines, voire éventuellement des milliers de milliards de paramètres) sont pré-entraînés pour la tâche, d'apparence bénigne, de prédire le mot suivant ou un mot masqué d'un texte. Leurs succès, même limités, dans des tâches pour lesquelles ils n'ont pas été prévus sont remarquables et en constante évolution. Néanmoins, il faut garder à l'esprit que ces systèmes sont construits sur la base de statistiques de proximité des mots dans les textes d'apprentissage. Ils sont pour le moment non factuels et incapables de raisonnement. Ils ne peuvent citer leurs sources de façon correcte ni vérifier la rationalité de leur propos. Ils peuvent énoncer des choses inexactes, qui semblent bien argumentées et peuvent avoir l'apparence de la réalité, mais qui peuvent aussi nuire aux personnes qu'ils nomment, en mentionnant par exemple des références inexactes et blessantes, voire injurieuses et diffamatoires. S'ajoute le fait que ces modèles soient entraînés à partir de données massives susceptibles d'être protégées par le droit des données personnelles, la propriété intellectuelle ou encore des règles deconfidentialité, tels les secrets d'affaires. La violation des lois n'est pas acceptable dans une société de droit et démocratique. Elle doit conduire à rechercher les moyens d'une responsabilité, à l'instar de ce que prévoient les institutions de l'Union européenne dans l'élaboration actuelle de la proposition de règlement sur l'IA (IA act). Notons qu'il s'agit là de problèmes inhérents à la génération actuelle de ces outils qu'il convient d'étudier avant de les rendre massivement utilisables spécialement dans des applications critiques. **Le respect de la règle de droit et des valeurs de notre société sont les conditions indispensables au développement d'une IA acceptable et de confiance.**

Les principes généraux de ces systèmes sont connus, dont l'apprentissage et l'ajustement fin par renforcement et retours humains afin de rendre les sorties socialement acceptables. Cependant de nombreux aspects pratiques de leur mise en œuvre relèvent de choix empiriques aux implications souvent complexes. Or, tous ces détails ne sont plus considérés comme tels lorsqu'il s'agit de mener une recherche rigoureuse. Ils revêtent une importance vitale pour s'assurer de la fiabilité, de l'innocuité sociale et éviter tout risque de biais et de discrimination. Ces objectifs doivent être poursuivis en parallèle de l'amélioration de leur performance et de leur robustesse, notamment face à des actes malveillants. Leur utilisation dans des systèmes critiques rend incontournable l'analyse de leur **explicabilité**, qui est une **fonction essentielle pour l'amélioration et la compréhension des décisions ou des arguments produits.** L'émergence d'aptitudes hors du domaine du traitement du langage naturel, touchant par exemple aux mathématiques, à l'informatique ou au sens commun, nécessite d'être expliquée et analysée, notamment en appui sur la logique. De plus, ces outils technologiques sont extrêmement énergivores lors de leur entraînement, ce qui rend leur mise au point inappropriée pour une utilisation fréquente dans la société. **Des efforts de recherche sont indispensables pour introduire des approches plus performantes** inspirées par exemple des neurosciences, afin de trouver des formulations aussi performantes, mais plus **frugales**. Cela constitue également un sujet prioritaire de la stratégie anationale portée par les 3IAs. Il s'avère également que certaines limitations de ces systèmes peuvent être repoussées grâce à leur couplage avec des modèles, des connaissances et d'autres systèmes d'IA. Dans cette optique, **l'apprentissage hybride, qui est au cœur d'ANITI, offre une voie de recherche que nous devons poursuivre et intensifier**.Toutes ces réflexions sur les problèmes de l'IA générative seront approfondies et alimenteront le programme de recherche et de formation, en cours d'élaboration, de la 2e phase d'ANITI. Il s'agira d'apporter une contribution forte vers une **IA de confiance, socialement utile et responsable, pour garantir son acceptabilité.**

CONTRIBUTEURS

Voici la liste des contributeurs et cosignataires, actifs dans le groupe de travail autour des IA génératiques et qui ont œuvré à la rédaction de cette note de synthèse.

Direction ANITI

Malik Ghallab : président Steering Commitee
Serge Gratton : directeur scientifique
Nicolas Viallet : directeur opérationnel
Corinne Joffre : secrétaire générale

Porteurs de chaire ANITI

Rachid Alami
Leïla Amgoud
Nicholas Asher
Jérôme Bolte
Jean-François Bonnefon
Céline Castets-Renard
Frédéric Dehais
Daniel Delahaye
Nicolas Dobigeon
Hélène Fargier
Fabrice Gamboa
Cesar Hidalgo
Jean-Bernard Lasserre
Bruno Jullien
Jean-Michel Loubès
Nicolas Mansard
Claire Pagetti
Jérôme Renault
Thomas Schiex
Thomas Serre
Joao Marques Silva
Sylvie Thiébaux
Louise Travé-Massuyès
Rufin VanRullen